Constructing a Syndromic Terminology Resource for Veterinary Text Mining

Furrer, Lenz ; Küker, Susanne ; Berezowski, John ; Posthaus, Horst ; Vial, Flavie ; Rinaldi, Fabio

Abstract: Public health surveillance systems rely on the automated monitoring of large amounts of text. While building a text mining system for veterinary syndromic surveillance, we exploit automatic and semi-automatic methods for terminology construction at different stages. Our approaches include term extraction from free-text, grouping of term variants based on string similarity, and linking to an existing medical ontology.

Constructing a Syndromic Terminology Resource for Veterinary Text Mining

Lenz Furrer
Institute of Computational Linguistics
University of Zurich
lenz.furrer@uzh.ch

Susanne Küker
Veterinary Public Health Institute
University of Bern
susanne.kueker@vetsuisse.unibe.ch

John Berezowski
Department of Clinical Research and Veterinary Public Health
University of Bern
john.berezowski@vetsuisse.unibe.ch

Horst Posthaus
Institute of Animal Pathology
University of Bern
horst.posthaus@vetsuisse.unibe.ch

Flavie Vial
Veterinary Public Health Institute
University of Bern
flavie.vial@vetsuisse.unibe.ch

Fabio Rinaldi
Institute of Computational Linguistics
University of Zurich
fabio.rinaldi@uzh.ch

Abstract

Public health surveillance systems rely on the automated monitoring of large amounts of text. While building a text mining system for veterinary syndromic surveillance, we exploit automatic and semi-automatic methods for terminology construction at different stages. Our approaches include term extraction from free-text, grouping of term variants based on string similarity, and linking to an existing medical ontology.

1 Introduction

In the project Veterinary Pathology Text Mining, we are developing tools to exploit veterinary post-mortem data for epidemiological surveillance and early detection of animal diseases. This paper describes the work in progress on the construction of a veterinary terminology resource as a basis for a text mining tool to classify, with minimal human intervention, free-text veterinary reports with respect to multiple clinical syndromes that can be monitored.

In human medicine, text mining has been successfully applied to clinical records in many public health surveillance systems (Botsis et al., 2011; Steinberger et al., 2008; Brownstein et al., 2008; Wagner et al., 2004). The approaches range from hand-written rule-based systems to fully automated methods using machine learning. For example, Chapman et al. (2004) use heuristical keyword-driven as well as supervised machine learning techniques (Naïve-Bayes classifier) for detecting mentions of fever in free-text clinical records. Similarly, the BioCaster system (Collier et al., 2006; Collier et al., 2008) relies on a carefully constructed medical ontology combined with a Naïve-Bayes classifier as an input filter. Friedlin et al. (2008) use a regular-expression based term-extraction system to find positive and negative mentions of methicillin-resistant *Staphylococcus aureus* in culture reports. Hartley et al. (2010) give an overview of surveillance systems that mainly focus on world-wide monitoring of web sources, including news feeds and informal medical networks.

The text mining of veterinary reports faces additional challenges such as multiple species and a less controlled vocabulary (Smith-Akin et al., 2007; Santamaria and Zimmerman, 2011). Up to this point, approaches for classifying veterinary diagnostic data into syndromes for surveillance have been restricted to the use of rule-based classifiers (Dórea et al., 2013; Anholt et al., 2014). To build these classifiers, a group of experts manually creates a large set of rules. The rules are then used to classify veterinary diagnostic submissions into syndromes based on the presence or absence of specific words within various fields in the diagnostic submission data.

We propose to develop a process for using text mining methodologies (natural language processing) to efficiently extract relevant health information from veterinary diagnostic submission data with minimal human intervention. Given a sufficient amount of data (i. e. at least a few hundreds of manually classified reports), a machine

learning approach will allow us to directly classify these data into syndromes that can be monitored for surveillance.

As recognized in the Swiss Animal Health Strategy 2010+, methods for early disease detection, based on the increasing abundance of data on animal health stored in national databases, can contribute to valuable and highly efficient surveillance activities. Post-mortem data, available from pathology services, are often under-exploited. The main purpose of post-mortem investigations of food production animals is to provide information about the cause of disease or death with regard to treatment, and prevention options for the affected herd. Besides these major diagnoses, all additional pathological findings are also recorded as text and electronically archived as necropsy reports. In addition to the value of this information for veterinarians and farmers, systematic evaluation of necropsy data may be of value the early detection of spatio-temporal clusters of syndromes which may result from a new disease emerging into a population or from changing patterns of endemic diseases. As such, it has the potential to be of value for both nation-wide and international (veterinary) public health early-warning systems.

The rest of this paper is organized as follows: We present our efforts in constructing and exploiting a veterinary terminology resource in Section 2. Section 3 describes our work towards report classification in the context of building a surveillance tool. The next steps and further application scenarios are given in Section 4.

2 Terminology Construction

In the process of report classification, we have put a lot of effort in the construction of a terminology resource that suited our needs. The resulting term inventory is tailored to a very specific task. Still, the methods, insights and even the resource itself can be of use for other applications. Similar to the work by Rinaldi et al. (2002), we extracted a set of terms from a collection of raw text and used automatic methods to organize them into a hierarchical structure. Section 2.1 introduces the categories we used for classification. In Sections 2.2 and 2.3, we describe the steps that led to the construction of the term inventory. Sections 2.4 and 2.5 show how this resource can be automatically enhanced for a more general usage.

2.1 Syndrome and Diagnosis Classification

The work described here is based on post-mortem reports that were compiled by the Institute of Animal Pathology (ITPA) of the Vetsuisse faculty at the University of Bern. The data were entered into a database by veterinary pathologists between 2000 and 2011. We used a subset of approximately 9 000 report entries regarding pigs and cattle. The reports are written in German, with a small fraction (less than 3 %) in English and French.

For subsequent quantitative analysis, we classified all reports using two categorization levels. As a coarse-grained categorization, we annotated each report with the syndromic groups that were affected by a medical issue. Each report was assigned zero, one or more of 9 syndrome categories (gastro-intestinal, respiratory, urinary, cardio-vascular, lymphatic, musculo-skeletal, reproductive, neural, other). This categorization approximately meets the level of granularity found in other work (Dórea et al., 2013; Warns-Petit et al., 2010). For a finer-grained categorization of the reports, we additionally annotated post-mortem diagnoses mentioned (directly or implicitly) in the reports, such as *enteritis*, *lipidosis*, or *injuries from foreign bodies*. The set of diagnoses was not defined a priori, but continuously updated in the classification process. The final set comprised some 100 classes and is shown in Table 1. The diagnoses are modeled as subcategories of the syndromes. While some category names occur in more than one syndromic category, it does not mean that they are ambiguous, as they are triggered by different terms. For example, *atresia* is classified as a congenital abnormality of the gastro-intestinal system, whereas the *ventricular septal defect* is a congenital abnormality of the cardio-vascular system.

2.2 Term Normalization

The medical reports have a high number of surface variants per term. The variation is caused by inflection, inconsistent spelling and typographical errors. On a higher level, variation is increased by synonymy, i. e. the use of different terms for the same concept (e. g. *Lipidose/Verfettung* 'lipidosis'). From the perspective of the given text mining task, certain derivative forms can be considered synonymous variants as well (e. g. *Ulzeration* besides *Ulkus*).

We split the report texts into tokens, which we

gastro-intestinal	
abomasal ulcer	178
abomasitis	120
acidosis	41
cheilitis	6
cholangitis	84
colitis	368
congenital abnormality	21
dilatation	159
displaced abomasum	55
duodenitis	12
enteritis	2458
esophagitis	46
gastric ulcer	206
gastritis	121
glossitis	16
hepatitis	200
HIS	361
Hoflund syndrome	10
icterus	39
ileitis	60
invagination	59
jejunitis	30
lipidosis	93
obstipation	13
omasitis	18
pancreatitis	2
parasites	13
perforation	35
pharyngitis	9
proctitis	13
reticulitis	98
rumenitic ulcer	4
rumenitis	92
sialoadenitis	2
steatorrhea	105
stenosis	10
stomatitis	57
trauma	25
typhlitis	37
volvulus	479
respiratory	
bronchiolitis	256
bronchitis	466
broncho-pneumonia	1040
laryngitis	23
pharyngitis	1
pleuritis	40
pneumonia	769
rhinitis	11
rhinitis atrophicans	196
sinusitis	8
tracheitis	28
urinary	
congenital abnormality	1
cystitis	56
hydronephrosis	30
nephritis	416
renal degeneration	116
trauma	8
urolithiasis	75
cardio-vascular	
cardiomyopathia	46
congenital abnormality	84
endocarditis	179
epicarditis	62
heart degeneration	50
hydropericard	319
myocarditis	77
pericarditis	427
pleuritis	41
lymphatic	
lymph-adenopathy	245
splenitis	77
tonsillitis	88
musculo-skeletal	
arthritis	231
arthrosis	31
bone degeneration	17
callus	14
congenital abnormality	30
fracture	94
luxation	9
myodegen-eration	70
myopathy	56
osteochondrosis	30
osteomyelitis	136
polyarthritis	343
synovitis	45
tendinitis	25
tendovaginitis	18
reproductive	
abortion	642
congenital abnormality	8
dystocia	9
metritis	70
perforation	4
placentitis	126
retained placenta	3
uterine perforation	4
uterine torsion	3
vaginitis	8
neural	
congenital abnormality	5
encephalitis	116
hydrocephalus	18
intoxication	5
meningitis	226
myelitis	13
neural degeneration	54
neuropathy	78
other	
crushed	81
dermatitis	184
enterotoxemia	285
eye related	22
foreign body	118
hernia	73
hydrothorax	104
inanition	74
intoxication	95
iron deficiency	65
mastitis	66
neoplasia	68
otitis	20
perforation	257
peritonitis	866
pleuritis	643
pododermatitis	19
polyserositis	297
rumen drinker	33
sepsis	647
splenic torsion	18
umbilicus related	117

Table 1: The diagnoses used for classification, grouped by syndrome, with number of occurrences.

defined as consecutive runs of alphanumeric characters or hyphens. We then performed a series of normalization steps in order to reduce the number of term variants when compiling an index.

The bulk of the spelling variation stems from Latin/Greek-originated terms, such as *Zäkum* 'cecum'. Besides the German spelling (using the letters *ä, ö, z/k*), the Latin spelling is often used (*ae, oe, c*, respectively), and even combinations of the two are encountered. For the previous example, the following variants are present, among others: *Caecum, caecum, Cäcum, Cäkum, Zaecum*. We normalized the usage of these letters by replacing *ä* with *ae* and *ö* with *oe* unconditionally, while treating *c* differently based on its right context: before a front vowel it was replaced by *z*, before *h* and *k* it was kept as *c*, and in all other cases (including word-final position) we replaced it with *k*. The complexity of this rule is owed to the fact that this normalization is applied to all words, i. e. including originally German words like *Kinn/Zinn* 'chin'/'tin', which would be confused by an unconditional conflation of *c, k, z*. As a side effect, the normalization of German terms occasionally captured closely spelled English terms (which were not systematically gathered), such as *Enzephalitis/encephalitis*.

Subsequently, we removed inflectional suffixes using the NLTK[1] implementation of the "Snowball" stemmer for German (Porter, 1980). Stemming is the process of removing inflectional and (partially) derivational affixes, thus truncating words to their stems. For example, *minimally* and *minimize* are both reduced to *minim* in Porter's English stemmer, which is not a proper word, but nev-

[1]Natural Language Toolkit: `www.nltk.org`

variants	normalized form	explanation
Zäkumtorsion, Caecumtorsion	zaekumtorsion	*ä/ö/c/k/z* normalization
Kokzidiose, Coccidiose	kokzidios	
Aborts, Abort, Abortes, Aborte, Aborten	abort	stemming
perforierter Ulcus, perforierten Ulkus	perforiert ulkus	both
Kardiomyopathie, Cardiomyopathie, Kardiomyopathien	kardiomyopathi	

Table 2: Normalization examples.

ertheless a useful key for lumping together etymologically related words.

Stemming is based on orthographical regularities and uses only a minimal amount of lexical information. Although the method is not flawless – it may be prone to errors with very short and irregularly inflected words – it generally works well for languages with alphabetic script and has been successfully applied to many European languages. Using a stemmer, we were able to considerably reduce the number of inflectional/derivational variants. However, a number of inflectional forms were still missed by the stemmer – especially plural forms with Latin inflection, such as *Ulkus/Ulzera*, or *Enteritis/Enteritiden*, which are not covered by the stemming rules for general German grammar. The stemmer also failed to capture most of the spelling errors. Table 2 illustrates the conflation with examples.

2.3 Focus Terms

For the syndromic classification of the veterinary reports, we manually created a list of focus terms which served as indicators for the clinical syndromes and diagnoses. Starting from a frequency-ranked list of the words found in all of the reports (already grouped by their normalized form), we manually selected terms that were likely to indicate (positive) diagnoses in the reports. The list was refined by inspecting the reports that produced hits for the focus terms.

The focus terms typically consist of a single token, but we also allowed multi-word expressions. The terms are grouped by diagnosis. Thus, each diagnosis refers to a set of terms which either constitute a common name of the diagnosis or describe some of its aspects. For example, concerning injuries caused by foreign bodies, we consider *Draht* 'wire' and *Nagel* 'nail' as focus terms, even though these words only refer to the cause, but not to the injuries themselves.

As each focus term is represented by its normalized form, a number of variant forms is already matched, as described above. We aimed to additionally cover variants produced by misspellings as well as inflected forms not recognized by the stemmer. Using approximate string matching, we searched the reports for similar terms for each of the focus terms. We used the `simstring` tool (Okazaki and Tsujii, 2010) for retrieving similarly spelled terms among the entire text collection. Approximate matching is a difficult task, as it is hard in general to formally define similarity among (the orthographical representations of) words in a way consistent with human judgement. `simstring` measures similarity as a function of the number of shared n-grams (runs of n characters) in two words, which is only a rough approximation of the task. However, compared to other similarity measures – e. g. Levenshtein's edit distance[2] – it is considerably more efficient for retrieval in large amounts of text. In the inevitable trade-off of good precision and high recall, we strove for recall by choosing a low similarity threshold for retrieval. As expected, this resulted in a high number of hits, including many false positives, i. e. words with a high n-gram similarity score, that are not actually similar to the input term (e. g. *arthritis* and *arteritis*). Due to the limited number of focus terms it was feasible to manually clean the list of similar words.

Figure 1 illustrates how term variants were gathered around the concept of a diagnosis. A number of synonymous and hyponymous terms were added to a specific diagnosis by a human expert. These terms were used as seeds to automatically find more variants, such as inflectional and spelling variants as well as misspellings. Please note that the labeled edges are only added for illustration purposes – the relations between term

[2]For a study of agreement between human judgement and different similarity measures, see e. g. Efremova et al. (2014); for a general overview of similarity measures cf. Navarro (2001) and Christen (2006).

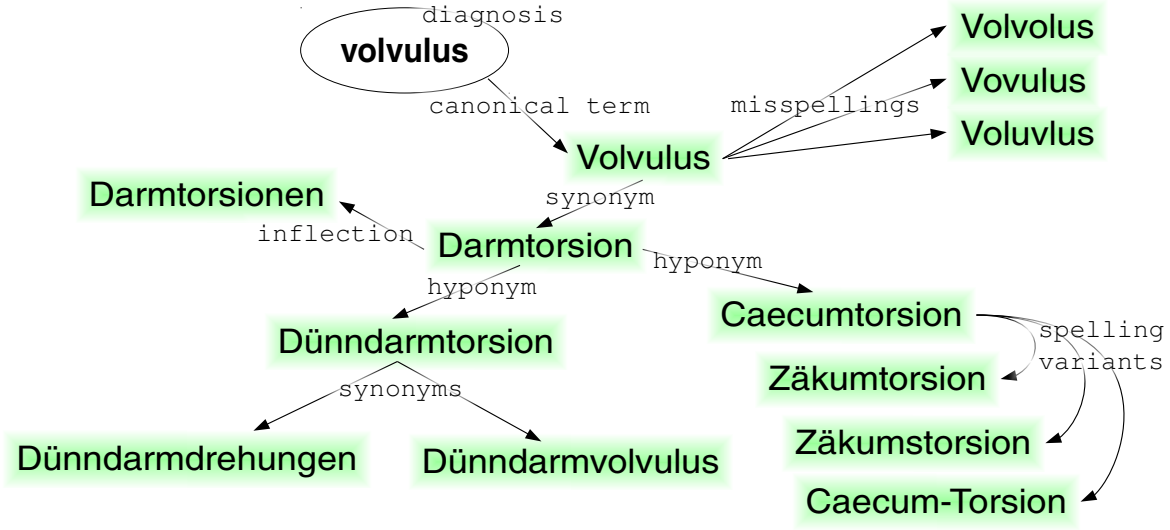

Figure 1: Term variants for the diagnosis *volvulus*.

forms (such as *synonym, misspelling*) were not captured during this phase, as they were not needed for syndrome/diagnosis classification. However, we examined ways to partly recover this underlying structure in an automated way, as is described in the following sections.

2.4 Further Term Conflation

The UMLS Metathesaurus[3] is a large collection of various medical terminology resources. One of its key features is the assignment of unique concept identifiers to entries from different vocabularies in many languages, thus establishing equivalence relations across them. By creating links to Metathesaurus concepts, we can enrich our own terminology resource with information contained in the Metathesaurus, as well as making it more valuable when sharing it with others.

We used the 2014AA release of the Metathesaurus for this work. For each concept that was represented in a German vocabulary, we normalized its lemma and tried to match it against an entry among our focus terms. With this approach, we were able to establish a link to one or more UMLS concepts for 80.6 % of the diagnoses.

Since our data were organized by diagnosis, each covering a number of terms with sometimes quite disparate meanings, the connection to the Metathesaurus produced a high number of one-to-many mappings (cf. Figure 2). This difference in

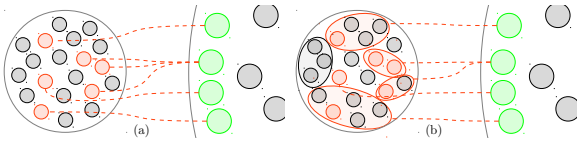

Figure 2: Ontology matching before (a) and after (b) term conflation. In both graphics, the left-hand side represents a diagnosis as a set of terms, some of which are linked to a UMLS concept (connected bullets) on the right-hand side.

granularity hinders the exploitation of the linked information, as the meaning of many diagnoses appears highly ambiguous in terms of the Metathesaurus. In order to better match the semantic range of the UMLS concepts, we passed on to perform the mapping at the level of terms rather than diagnoses. This required us to add a hierarchical layer to our data structure: We needed to distinguish *term variants* (spelling and inflectional alternations, such as *Caecumtorsion* vs. *Zäkumstorsion*) from *separate terms* (e. g. *Zäkumstorsion* vs. *Darmtorsion*). Please note that synonyms such as *Darmtorsion* and *Darmdrehung* are considered separate terms, even though they have the same meaning.

For each diagnosis, we organized all term forms into groups of term variants. The arrangement was performed automatically, based on string similarity. While string similarity is only an unreliable approximation of human similarity judgement, and while there are a number of concurring ways of computing it, it is also difficult to determine a

[3] www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

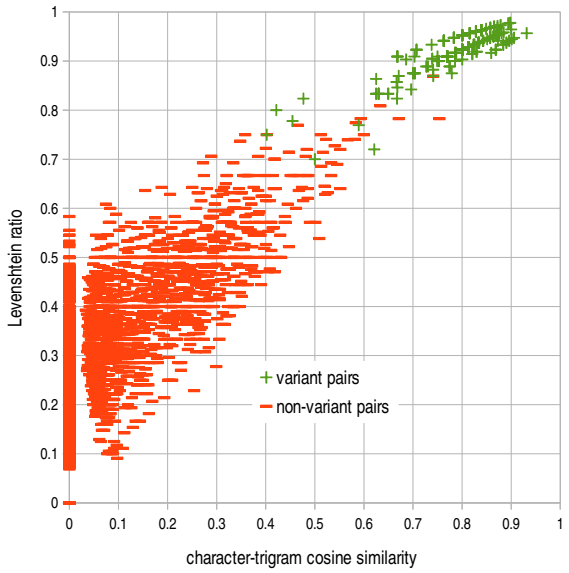

Figure 3: Two different similarity measures for pairs of similar and dissimilar words.

threshold that clearly separates similar from dissimilar pairs of words. We therefore chose to perform supervised machine learning, i. e. automatic learning by example. We compiled a training set of positive instances of inflectional/spelling alternation as well as negative instances, i. e. pairs of unrelated words. For each pair, we computed two different string similarity measures (cf. Figure 3): cosine similarity of character trigram vectors, and Levenshtein ratio. These two measures cover different aspects of similarity, and thus their combination might capture more information than just one of them. We trained a Support Vector Machine on the two-dimensional space of the similarity measures, using a polynomial kernel function.

The automatic term grouping yielded very satisfactory results. We manually evaluated the resulting groups, requiring that all members be orthographical or inflectional variations of each other. We also allowed derivational variants (e. g. *Weiss-muskelkrankheit/…erkrankung* 'white muscle disease') to be in the same group, although the separation of derivatives (e. g. *Ulkus/Ulzeration*) was not counted as false negative. We found that less than 6.7 % of the groups contained unequal terms (false positives), and only 1.9 % of the groups were erroneously isolated instead of being merged with the correct equivalents (false negatives). Many false positive judgements were caused by terms with only small differences in meaning, such as *Muskeldegeneration* 'muscle degeneration' and *Muskelfaserdegeneration* 'muscle fiber degeneration', which might even be regarded equal in a less strict evaluation. As for the false negatives, the number of misses could be reduced by extending the stemmer with Latin-inflection endings like *Ulkus – Ulzera*.

2.5 Connecting to UMLS

Each group of term variants was then linked to a UMLS concept if there was a match between at least one member of the group (i. e. a term variant) and of the German concept descriptions, respectively. Only exact agreement of the normalized forms was counted as a match, as preliminary experiments had shown that fuzzy matching introduced a great amount of false positives (connections between similarly spelled, but otherwise unrelated words) while adding only very few desired links. However, we were able to improve the linkage with simple heuristics, such as the removal of boilerplate expressions like *nicht näher bezeichnet* 'not otherwise specified'.

In 42.1 % of the terms, we could find a match with a UMLS concept. Only 6.7 % of the matching terms point to more than one concept, which means that 93.3 % of the terms with a match can be mapped to the Metathesaurus unambiguously. However, for more than half of the terms no corresponding UMLS concept could be found at all, which is mainly due to the different domains of our veterinary texts and the predominantly human-medicine-based UMLS. Table 3 shows some examples of the mapping.

The connections to the Metathesaurus allowed us to further enrich our data. For example, every UMLS concept has a semantic type assigned to it, such as "Disease or Syndrome" or "Pathologic Function". Additionally, we used the concept descriptions in Metathesaurus to find more focus terms. By matching the descriptions of connected concepts against our text collection, we were able to enlarge the set of focus terms by almost 10 %.

As next steps, we plan to create links to other widely-used terminology resources, such as the *Central key for health data recording* by the International Committee for Animal Recording (ICAR).[4]

[4] See `www.icar.org`

diagnosis/terms	UMLS
stenosis	
Darmstenose	C0267465 Darmstenose/Darmstriktur/Stenose des…
Dünndarmstenose	C0151924 Dünndarmstenose/Stenose des Dünndarms
Rectumstenose, Rektumstenose	–
myodegeneration	
Belastungsmyopathie	–
Muskelfaserdegeneration, Muskelfaserndegeneration	C0234958 Muskeldegeneration/Degeneration des …
Muskelfasernekrose	–
Muskelläsionen	–
Muskelnekrose, Muskelnekrosen	C0235957 Muskelnekrose/Myonekrose
Myodegeneration, myodegeneration	–
Myonekrose, myonecrosis	C0235957 Muskelnekrose/Myonekrose
Rhabdomyolyse	C0035410 Rhabdomyolyse
Weiss-Muskel-Krankheit, Weiss-Muskelkrankheit, Weissmuskelerkrankung, Weissmuskelkrankheit, Weissmuskelkrankheit	C0043153 Muskeldystrophie, nutritive/ Weißmuskelkrankheit

Table 3: Mapping to the UMLS Metathesaurus.

3 Annotation Tool

The terminology resource described above is a key component in our efforts to create a veterinary surveillance system. We wrote a pipeline of Python scripts that assists our semi-automatic annotation of the pathology reports. The tool performs automatic annotation of syndromes and diagnoses based on the term resource, while also keeping track of manual verfications and rejections. Through a web interface, it accepts a Microsoft Excel workbook as input and produces a modified version in the same format, which allows a veterinary domain expert to inspect and modify the automatic annotations. All relevant information – such as the term resource and the assigned categories, negations (see below), and the previous manual annotations – are contained within this file.

3.1 Negation Detection

In a keyword-based system for detecting evidence, negative expressions can play a crucial role. Occasionally, negative outcomes of an analysis are reported in the texts, and suspected diagnoses are rejected quite frequently, such as *keine Hinweise auf eine Pneumonie* 'no evidence of a pneumonia'. Therefore, we aimed at identifying occurrences of focus terms that are mentioned in a negated context.

Besides the identification of negated expressions, negation detection heavily depends on the correct determination of their scope. Tanushi et al. (2013) compare different approaches to negation scope detection in Swedish clinical reports. According to them, "[e]mploying a simple, rule-based approach with a small amount of negation triggers and a fixed context window for determining scope is very efficient and useful, if results around 80 % F-score are sufficient for a given purpose" (Tanushi et al., 2013, p. 393). We included a simple negation-detection module in our pipeline, which looks for a set of negative expressions in a context window of 5 tokens to either side of the focus term. The context can be restricted for each expression (e. g. only to the right of or only immediately preceding a focus term). The context window is shortened at sentence boundaries and other indicators of a break. However, as the results of the negation detection are not yet satisfactory, we plan to integrate an existing library for this task, e. g. the Python package pyConTextNLP (Chapman et al., 2011).

3.2 Inter-Annotator Agreement

In order to validate the quality of our annotations, we organized a multi-annotator evaluation. We performed an experiment with six experts of veterinary pathology, which were asked to classify a number of reports with respect to the syndromic categories described in Section 2.1. For this purpose, we created a web interface which displayed the report text together with some metadata, one report at a time, and allowed to mark each of the syndromes as present or absent. The reports were randomly sampled, keeping the distribution

syndrome	reports		D_o	D_e	α
gastro-int.	52	(13)	0.059	0.251	**0.764**
respiratory	28	(10)	0.045	0.207	**0.781**
urinary	9	(3)	0.014	0.083	**0.836**
cardio-vasc.	15	(9)	0.041	0.115	**0.644**
lymphatic	3	(3)	0.014	0.018	**0.240**
musc.-skel.	13	(3)	0.014	0.125	**0.891**
reproductive	9	(1)	0.005	0.094	**0.952**
neural	5	(2)	0.009	0.052	**0.825**
other	38	(21)	0.095	0.226	**0.577**
avg.					**0.723**

Table 4: Inter-annotator agreement of the syndromic categories, measured with Krippendorff's Alpha. The second column gives the number of reports where at least one annotator marked the corresponding syndrome as present; following in parentheses is the number of reports with disagreement. D_o and D_e are the observed and expected disagreement, respectively.

of species and year of creation as close to the entire collection as possible (approaching stratified sampling). Each annotator was provided with a sample of 20 reports, which was extended to twice or three times the size when an annotator asked for more. In order to increase sample size, the same report was given to only two or three annotators, rather than all of them. In total, 81 distinct reports were annotated.

We evaluated the inter-annotator agreement with Krippendorff's Alpha (Krippendorff, 2013, pp. 267–309), as is shown in Table 4. For computing the agreement, we regarded each syndrome as an independent, binary variable (each syndrome is either present or absent in a report). The agreement value α ranges from 1 (perfect agreement) to 0 (agreement as by chance) or even below (systematic disagreement). A high agreement means that identifying syndromes is a clear task, while a low agreement indicates that the decisions cannot be easily made. Most of the syndromes have a good (>0.8) or acceptable (>0.6) α score,[5] whereas some are clearly identified as problematic. For the lymphatic system, the sparse representation (only 3 reports) does not allow for valid conclusions; further investigation is required in this case. The "catch-all" class *other*, however, most likely suffers from having an unclear scope. As a consequence of this evaluation, we decided to reduce the ambiguity of *other* by including additional classes in the next revision of the syndromic categorization.

[5]For a discussion of the interpretation of absolute agreement scores see Artstein and Poesio (2008, p. 591)

4 Outlook

We will assess the performance of the text-mining tool based on a small number of diseases which have been relevant in Switzerland in the last 10 years:

1. Bovine Viral Diarrhoea in cattle (an eradication campaign for the disease was introduced in 2008)
2. Porcine Circovirus type 2 infection in pigs
3. Gastro-intestinal syndromes in pigs (for which we observe an increasing amount of pathology submissions)

Time-series analyses will be performed to quantify trends, seasonality and other effects (day of week, day of month etc.) on the number of submissions for syndromes potentially related to these diseases. For each disease, "in-control" data (data collected in the absence of an outbreak) will be used to establish a baseline model describing the amount of normal "noise" in the data (expected number of submissions in the absence of disease outbreaks). Retrospective analyses of the time-series will be done to see whether alerts (signals) were produced when the number of submissions for syndromes potentially linked to the disease was higher than expected from our baseline model (event detection). This will allow us to evaluate whether the system would have worked as an early-warning system.

The tools developed in this project will be adapted to reports from different pathology institutes throughout Switzerland, thus contributing to a nation-wide syndromic surveillance system. Similarly, the methodology developed may be applicable to the analysis of text-based disease information which is recorded in other contexts. For example, there is a great potential of using such a system to systematically analyse health data which are recorded by veterinary practitioners in their practice management software, slaughter data or by animal health services in their central database.

Acknowledgements

This work was funded by the Swiss Federal Food Safety and Veterinary Office (Bundesamt für Lebensmittelsicherheit und Veterinärwesen).

References

R. Michele Anholt, John Berezowski, Iqbal Jamal, Carl Ribble, and Craig Stephen. 2014. Mining free-text medical records for companion animal enteric syndrome surveillance. *Preventive Veterinary Medicine*, 113(4):417–422.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Taxiarchis Botsis, Michael D. Nguyen, Emily Jane Woo, Marianthi Markatou, and Robert Ball. 2011. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *Journal of the American Medical Informatics Association*, 18(5):631–638.

John S. Brownstein, Clark C. Freifeld, Ben Y. Reis, and Kenneth D. Mandl. 2008. Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med*, 5(7):e151.

Wendy W. Chapman, John N. Dowling, and Michael M. Wagner. 2004. Fever detection from free-text clinical records for biosurveillance. *Journal of Biomedical Informatics*, 37(2):120–127.

Brian E. Chapman, Sean Lee, Hyunseok Peter Kang, and Wendy W. Chapman. 2011. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *Journal of Biomedical Informatics*, 44(5):728–737.

Peter Christen. 2006. A comparison of personal name matching: Techniques and practical issues. Technical Report TR-CS-06-02, The Australian National University, Dec.

Nigel Collier, Ai Kawazoe, Lihua Jin, Mika Shigematsu, Dinh Dien, Roberto A. Barrero, Koichi Takeuchi, and Asanee Kawtrakul. 2006. A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language Resources and Evaluation*, 40(3-4):405–413.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, et al. 2008. BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

Fernanda C. Dórea, C. Anne Muckle, David Kelton, J. T. McClure, Beverly J. McEwen, W. Bruce McNab, Javier Sanchez, and Crawford W. Revie. 2013. Exploratory analysis of methods for automated classification of laboratory test orders into syndromic groups in veterinary medicine. *PLOS one*, 8(3):e57334.

Julia Efremova, Bijan Ranjbar-Sahraei, and Toon Calders. 2014. A hybrid disambiguation measure for inaccurate cultural heritage data. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, page 47–55, Gothenburg, Sweden, April. Association for Computational Linguistics.

Jeff Friedlin, Shaun Grannis, and J. Marc Overhage. 2008. Using natural language processing to improve accuracy of automated notifiable disease reporting. *AMIA Annual Symposium Proceedings*, 2008:207–211.

David Hartley, Noele Nelson, Ronald Walters, Ray Arthur, Roman Yangarber, Larry Madoff, Jens Linge, Abla Mawudeku, Nigel Collier, John Brownstein, Germain Thinus, and Nigel Lightfoot. 2010. Landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, 3(e3).

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Thousand Oaks, CA, 3rd edition.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, March.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, page 851–859, Beijing, China, August.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Fabio Rinaldi, James Dowdall, Michael Hess, Kaarel Kaljurand, Mare Koitand, Kadri Vider, and Neeme Kahusk. 2002. Terminology as knowledge in answer extraction. In *TKE-2002: 6th International Conference on Terminology and Knowledge Engineering*, Nancy, France, August.

Suzanne L. Santamaria and Kurt L. Zimmerman. 2011. Uses of informatics to solve real world problems in veterinary medicine. *Journal of veterinary medical education*, 38(2):103–109.

Kimberly A. Smith-Akin, Charles F. Bearden, Stephen T. Pittenger, and Elmer V. Bernstam. 2007. Toward a veterinary informatics research agenda: An analysis of the PubMed-indexed literature. *International Journal of Medical Informatics*, 76(4):306–312.

Ralf Steinberger, Flavio Fuart, Erik van der Goot, Clive Best, Peter von Etter, and Roman Yangarber. 2008. Text mining from the web for medical intelligence. In Françoise Fogelman-Soulié, Domenico Perrotta, Jakub Piskorski, and Ralf Steinberger, editors, *Mining Massive Data Sets for Security*, volume 19 of *NATO Science for Peace and Security Series – D: Information and Communication Security*, page 295–310. IOS Press.

Hideyuki Tanushi, Hercules Dalianis, Martin Duneld, Maria Kvist, Maria Skeppstedt, and Sumithra Velupillai. 2013. Negation scope delimitation in clinical text using three approaches: NegEx, PyConTextNLP and SynNeg. In *Proceedings of the*

19th Nordic Conference of Computational Linguistics (NODALIDA 2013), page 387–397, Oslo, Norway.

Michael M. Wagner, J. Espino, F-C. Tsui, P. Gesteland, W. Chapman, O. Ivanov, A. Moore, W. Wong, J. Dowling, and J. Hutman. 2004. Syndrome and outbreak detection using chief-complaint data – experience of the Real-Time Outbreak and Disease Surveillance project. *Morbidity and Mortality Weekly Report*, 53:28–31.

Eva Warns-Petit, Eric Morignat, Marc Artois, and Didier Calavas. 2010. Unsupervised clustering of wildlife necropsy data for syndromic surveillance. *BMC Veterinary Research*, 6(56):1–11.